

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan *Artificial Intelligence* bertumbuh secara pesat dalam bidang riset maupun penerapannya di dunia nyata. Salah satu perkembangan yang pesat terlihat pada komputasi linguistik, contohnya adalah pemrosesan bahasa alami (*Natural Language Processing*). Menurut *MarketsandMarkets* perkembangan pasar NLP diperkirakan akan meningkat dari 7.63 miliar dolar pada tahun 2016 menjadi 16.07 miliar dolar (Brunner, 2018). Angka tersebut menandakan implementasi NLP yang signifikan dalam dunia nyata. Dalam bidang riset perkembangannya ditandai dengan mulai digunakan algoritma dan metode baru untuk meningkatkan akurasi dari model pemrosesan bahasa alami. Model – model *machine learning* seperti *Recurrent Neural Network*, *Convolutional Neural Network* dalam beberapa implementasinya terbukti meningkatkan tingkat akurasi model NLP.

Kemudahan untuk mendapat data linguistik adalah salah satu aspek penting yang mendorong populernya penggunaan NLP. Semenjak perkembangan internet, teks digital dapat lebih mudah dicari. Namun masih ada batasan yang besar dalam perkembangan NLP, salah satunya adalah sistem dan sumber daya NLP yang umum sekarang hanya tersedia dalam *High-resource Language (HRL)* atau Bahasa yang memiliki sumber daya besar seperti Bahasa Inggris, Perancis, Spanyol, Jerman, Cina. Bahasa – Bahasa yang tergolong *low-resource Language (LRL)* seperti Bahasa Indonesia, Punjabi, Swahili maupun bahasa lain yang digunakan banyak orang masih belum memiliki sumber daya atau sistem yang cukup (Hirschberg & Manning, 2015). Model *HRL* yang sudah ada tidak bisa langsung diimplementasikan untuk mengidentifikasi struktur kalimat dari bahasa yang

tergolong *LRL*, karena terbatasnya *dataset* yang tersedia. Selain terbatasnya *dataset*, model *HRL* yang tersedia dioptimisasi untuk tugas spesifik dari Bahasa *HRL* sehingga akan mempengaruhi kemampuan model untuk memprediksi Bahasa *LRL*.

Berdasarkan masalah diatas, penelitian ini akan merancang dan membangun sebuah model NLP untuk mengidentifikasi struktur kalimat Bahasa Indonesia yang dikembangkan dengan bahasa pemrograman *Python*. Selain perancangan model, dikumpulkan juga *dataset* Bahasa Indonesia untuk keperluan *training* dan evaluasi model NLP. Melalui model yang dibuat diharapkan dapat menambah sistem dan sumber daya NLP Bahasa Indonesia yang masih terbatas. Model yang dibuat juga diharapkan dapat mengejar ketertinggalan perkembangan Bahasa Indonesia dalam bidang pemrosesan bahasa alami.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas maka rumusan masalah adalah sebagai berikut:

1. Bagaimana membangun sebuah model pembelajaran mesin untuk mengidentifikasi struktur kalimat Bahasa Indonesia dengan *Python*.
2. Bagaimana menganalisis kemampuan model untuk memprediksi *tag IOB* dari struktur kalimat Bahasa Indonesia.

1.3 Batasan Masalah

Adapun yang menjadi batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Bahasa Pemrograman yang digunakan adalah *Python*.
2. Model yang dibuat hanya digunakan untuk mengidentifikasi *tag IOB* struktur kalimat Bahasa Indonesia.
3. Model pembelajaran mesin yang digunakan adalah LSTM dan BiLSTM.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah:

- 1 . Membangun sebuah model pembelajaran mesin untuk identifikasi struktur kalimat Bahasa Indonesia.
- 2 . Menganalisis kemampuan model untuk memprediksi *tag IOB* dari struktur kalimat Bahasa Indonesia.

1.5 Metodologi Penelitian

Metode yang digunakan dalam penelitian ini adalah:

1. Metode Studi Pustaka/Literatur

Metode penelitian kepustakaan digunakan untuk mencari literatur atau sumber pustaka yang berkaitan dengan analisis model *Natural Language Processing* yang pernah dibuat dan diterapkan.

2. Metode Pengumpulan Data

Dalam penelitian ini, data dikumpulkan dengan cara mengambil kalimat yang memenuhi standar S-P-O-K dari berbagai sumber.

3. Perancangan Model

Dalam penelitian ini, metode perancangan model dilakukan dengan merancang model pembelajaran mesin, dan fungsi-fungsinya.

4. Pengkodean Model

Proses pembangunan program yang merealisasikan rancangan model yang dikembangkan dengan menggunakan Bahasa pemrograman Python.

5. Pengujian Model

Proses pengujian terhadap model yang dibuat, apakah model memiliki akurasi yang cukup untuk identifikasi struktur kalimat. Proses pengujian meliputi *training* dan *testing* model pada dataset Bahasa

Indonesia yang telah dikumpulkan.

